William Eduardo Soto Martinez

Nancy, France | +33 07 54 13 00 59 | williamsotomartinez@gmail.com
Website: sotwi.github.com | LinkedIn: Wilsoto

Profile

Multilingual NLP researcher with 5+ years of experience specializing in multilingual graph-to-text generation and evaluation across high- and low-resource languages. Published on top conferences (ACL, INLG, IJCNLP-AACL) with research on RDF, AMR, Soft Prompting, and QLoRA. Skilled in Python, PyTorch, and Transformers, with a strong track record of collaborative research and practical ML development. Committed to advancing inclusive, data-efficient language technologies.

Skills

Technical Skills

- **Programming Languages:** Python (Advanced), Java (Intermediate), C++ (Intermediate)
- **Machine Learning & NLP:** PyTorch (Advanced), Transformers (Advanced), SpaCy (Advanced)
- **Data Management:** Pandas (Advanced), Sparql (Intermediate), SQL (Intermediate)
- **Deployment Tools:** Git (Advanced), Anvil (Advanced), Gradio (Intermediate)

Additional Competencies

- **Model Fine-Tuning:** Text generation and evaluation.
- **PEFT techniques:** Soft Prompting and LoRA.

Soft Skills

- Collaborative Research
- Cross-Cultural Communication
- Mentoring and Teaching

Languages:

Spanish (C2), English (C1), French (B1)

Professional Experience

Centre National de la Recherche Scientifique — **Nancy, France**

Research Engineer - Multilingual KG-to-Text Evaluation — **Oct. 2024 - Oct. 2025**

- Engineered and fine-tuned evaluation models on 1.7 million samples covering six languages (tested on seven, including one unseen), utilizing Transformer architectures (mDeBERTa) and PEFT techniques (LoRA). Accepted to ACL 2025.
- Achieved 5 to 10% improvement in correlation with indirect human evaluations and demonstrated moderate-to-strong alignment with direct human judgments; deployed as a Hugging Face space.
- Supervised an intern to run prompting experiments for a paper currently under review for EMNLP 2025.

Doctoral Researcher - Multilingual Graph-to-Text Generation — **Oct. 2021 - Oct. 2024**

- Led development of multilingual graph-to-text pipelines using MT5 PEFT techniques (soft prompts, LoRAs). Trained AMR-to-Text on 30000 samples per language across 12 languages (published on INLG 2024) and RDF-to-Text on 1500 samples per language across 4 languages (on IJCNLP-AACL 2023).
- Achieved state-of-the-art performance in Romance languages and matched top systems in Germanic languages with at least 50% less data, added support for six previously unsupported languages, and deployed to Hugging Face spaces.
- Instructed around 30 students of an MSc in NLP as well as other students at the Python for NLP Summer School 2020, consistently earning top feedback for clarity and engagement.

Master 2 Research Intern - Multilingual Paraphrase Evaluation **Mar. 2021 - Aug. 2021**

- Designed and implemented an end-to-end pipeline for paraphrase evaluation across multiple languages using custom Python scripts and Transformer-based metrics.
- Streamlined evaluation workflows, enabling fast analysis and informing improved paraphrase generation.

Master 1 Research Intern - Language Identification **Jun. 2020 - Jul. 2020**

- Compiled a Guadalupean Creole corpus of more than 4000 sentences from multiple sources, and used it to further fine-tuned an existing FastText based LID classifier (presented on LIFT 2020).
- Reached 90% F1 score on the previously unsupported Guadalupean Creole language.

Universidad de Costa Rica **San Pedro, Costa Rica**

Research and Teaching Assistant **Mar. 2017 - Aug. 2019**

- Conducted data anonymization, multi-agent simulation, and cloud computing using Python and Logo.
- Supported BSc-level courses in compilers, data bases, and computational paradigms.

Education

Université de Lorraine **Nancy, France**

PhD in Informatics **Defense on Oct. 2025**

Thesis subject: Multilingual Graph-to-Text Generation and Evaluation

MSc in Natural Language Processing **2021**

Thesis subject: X-ParEval: A Multilingual Metric for Paraphrase Evaluation.
Grade Average: 16.4/20
Mention: Très Bien

Universidad de Costa Rica **San Pedro, Costa Rica**

BScn in Computer Sciences **2017**

Grade Average: 8.7/10

Publications

- Semantic Evaluation of Multilingual Data-to-Text Generation via NLI Fine-Tuning: Precision, Recall and F1 scores (Soto Martinez et al, ACL 2025) [Git] [Demo]
- Generating from AMRs into High and Low-Resource Languages using Phylogenetic Knowledge and Hierarchical QLoRA Training (HQL) (Soto Martinez et al, INLG 2024) [Git] [Demo]
- Phylogeny-Inspired Soft Prompts For Data-to-Text Generation in Low-Resource Languages (Soto Martinez et al, IJCNLP-AACL 2023) [Git]
- Language Identification of Guadeloupean Creole (Soto Martinez, LIFT 2020) [Git]

Collaborations

- NL-Augmenter: A Framework for Task-Sensitive Natural Language Augmentation (Dhole et al., NEJLT 2023) [Git]
- The 2023 WebNLG Shared Task on Low Resource Languages. Overview and Evaluation Results (WebNLG 2023) (Cripwell et al., MMNLG 2023)

Repositories

Github: Sotwi | Inria Gitlab: Wsotomar | Gitlab: Williamsotomartinez | Hugging Face Hub: WilliamSotoM

Interests

Languages, Literature, Reading, Creative Writing, Board Games, Game Development, Stargazing.